ANÁLISIS DE ALGORITMOS PARA EL AGRUPAMIENTO DE MUESTRAS METAGENÓMICAS

ADRIANA MARÍA ESCOBAR VASCO

Trabajo de grado para optar al título de grado

Isis Bonet Cruz Ph.D

ESCUELA DE INGENIERÍA DE ANTIOQUIA
INGENIERÍA DE SISTEMAS Y COMPUTACIÓN
ENVIGADO
2015

CONTENIDO

pág.

LISTA DE TABLAS

pág.

LISTA DE FIGURAS

pág.

LISTA DE ANEXOS

pág.

RESUMEN

Las formas de vida microscópicas se encuentran en todos los lugares y ambientes del planeta, y en su genética se halla información de gran valor para los científicos, sin embargo al tomar una muestra para estudiarlos solo se logra aislar y cultivar menos del 1% de ellos. La metagenómica nace con el fin de estudiar el otro 99% de la muestra y de descubrir más acerca de estas comunidades. El objetivo de la metagenómica es la secuenciación y el análisis de los genes contenidos en los cromosomas de microorganismos, esfuerzos en los cuales se enfoca el presente proyecto.

Para la realización del proyecto se utiliza el programa weka y el algoritmo k-means, implementado en una versión iterativa que utiliza la distancia coseno o euclidiana dependiendo del criterio del experto. Además utiliza cómo parámetro las distancias entre los clústeres para escoger los mejores y optimizar los resultados de la segunda iteración.

Con en desarrollo de este proyecto se llega a la conclusión que el k-means iterativo es una mejora al k-means, optimizando los resultados y encontrando clústeres más puros. También se encuentra que los resultados obtenidos con la función de distancia coseno son mejores que aquellos obtenidos con la función euclidiana y que el mejor atributo para describir las distancias es el 4-mer.

Palabras clave: metagenómica, k-means, clusterización, Weka.

ABSTRACT

Microscopic life forms exist in every place and environment in this planet. In their genetics there is valuable information for scientists, but only 1% of these microorganisms can be separated from a sample and cultivated to be studied. Metagenomics is the name of the field that is focusing on the study of the other 99% of the microorganisms in the sample. Its main goal is to sequence and analyze the genes in the chromosomes of the microorganisms, and learn as much as possible about the microbial world. This project emphases on the task of grouping the sequences of genes and identifying them.

The program Weka and the algorithm k-means are used for the development of this project. An iterative version that provides the Euclidian and cosine distance functions as a parameter for the user to choose is implemented. The distance between the clusters is used to choose the best groups and optimize the results of the second iteration.

As a result of this project a clear difference between the iterative k-means and the original one is found. This new algorithm is an optimization of the older one finding clusters that are more pure. The conclusion that cosine is a better distance function to calculate the clusters was reached, and 4-mer as an attribute to describe instances is the best choice.

Keywords: metagenomics, k-means, clusterization, Weka

INTRODUCCIÓN

Las formas de vida microscópicas dominan el planeta desde diferentes puntos de vista, no sólo porque existen en grandes cantidades, sino también porque son capaces de subsistir en gran variedad de ambientes y condiciones, y son clave para mantener otras formas de vida. Los microorganismos son los responsables de epidemias en humanos, animales y cultivos, pero a su vez traen consigo grandes beneficios para la vida humana en el área de alimentación, agricultura, medicina, entre muchas otras que se ven afectadas por los microorganismos (Wooley JC, 2010). Es por esto que los científicos han sentido gran curiosidad por los diferentes microorganismos existentes, y es justo con este objetivo que nace la metagenómica.

El término metagenómica se refiere a la secuenciación y al análisis de genomas de microorganismos, es decir, el análisis de los genes contenidos en sus cromosomas (Kislyuk, Bhatnagar, Dushoff, & Weitz, 2009). La metagenómica enfoca sus esfuerzos en estudiar organismos que no se pueden cultivar fácilmente en un laboratorio; esto hace que las secuencias de ADN utilizadas sean tomadas directamente del medio natural (Diaz, Krause, Goesmann, Niehaus, & Nattkemper, 2009). Esta área ha experimentado un alto crecimiento, no solo debido a la cantidad de organismos que cumplen con la característica previamente mencionada, sino también porque es de gran importancia comprender el contenido genómico de los organismos para conocer cuáles son sus roles y sus interacciones en un ecosistema determinado (Li, Wooley, & Godzik, 2008).

En el análisis de la metagenómica se encuentran dos grandes retos, el primero surge ya que los datos recolectados se encuentran fragmentados, y no se sabe con certeza a que organismos pertenece cada fragmento, aún más no se conoce con exactitud que organismos se encuentran en la muestra tomada. El segundo problema con el cual se enfrentan está relacionado con la extensa magnitud de las bases de datos biológicas que se deben analizar; para reconocer los fragmentos tomados, e identificar de cuál organismo hacen parte se deben comparar los fragmentos recolectados con las bases de datos existentes. Este proceso requiere gran capacidad computacional.

Para identificar la pertenencia de una secuencia de ADN a un clúster se utilizan principalmente, dos grandes métodos. Uno de estos métodos está basado en la composición genómica de las secuencias, evaluando factores como el promedio de contenido de las bases nitrogenadas Guanina y Citosina; con esta información se obtiene sugerencias sobre las características que presenta un organismo específico, además tiene las ventajas de ser un proceso automático y rápido, que puede ser utilizado tanto en clasificación como en clusterización (McHardy, Martin, Tsirigos, Hugenholtz, & Rigoutsos, 2007). El segundo método se basa en la similitud que presentan las secuencias estudiadas, este método utiliza alineación de secuencias para encontrar las equivalencias entre los diferentes fragmentos obtenidos, a diferencia del primero que se basa en patrones estadísticos de la distribución de los oligonucleótidos (Kislyuk et al., 2009). Un ejemplo de un programa basado en composición es BLAST.

El programa BLAST tiene gran popularidad en el medio y es utilizado en la mayor parte de los estudios realizados. Este permite la comparación de una secuencia con una base de datos (Diaz et al., 2009). Para lograrlo utiliza un algoritmo heurístico que busca alineamientos locales, detectando relaciones entre las secuencias que comparten regiones similares (Li et al., 2008). Este programa demanda grandes cantidades de tiempo, por lo que se busca optimizar el proceso mejorando la calidad de la información que se ingresa en el programa.

Los métodos de creación de grupos o clústeres conocidos como algoritmos de clusterización, son comúnmente utilizados para el análisis rápido de la diversidad de las secuencias y de la estructura interna de la muestra. Esto lo logran agrupando secuencias similares en clústeres, identificando familias presentes en la muestra, y como resultado facilitan la comparación de genomas (McHardy et al., 2007). Además de este método de agrupación basado en la similitud de los genomas, existe uno basado en la composición de las secuencias de ADN que se enfoca en características de este como lo son las frecuencias de oligonucleótidos, los cuales no solo dan información de la estructura de ADN, sino también tienen la ventaja que son constantes en un genoma y permiten predecir la phylogenetica, esta es una característica del ADN que ayuda a inferir la relación existente entre grupos de organismos facilitando así la creación de grupos (Kelley & Salzberg, 2010).

El objetivo del presente trabajo es encontrar un método óptimo para el preprocesamiento de esta información.

1. PRELIMINARES

1.1 PLANTEAMIENTO DEL PROBLEMA

Debido al interés que existe en el área de la metagenómica se ve la importancia de analizar las secuencias genómicas encontradas en las muestras e identificar a que organismo pertenece cada una. Hoy en día existen dos métodos que buscan resolver esta situación, uno basado en la composición de los genes, el cual no es totalmente preciso y conlleva tiempo computacional; el otro se basa en similitud que presentan las secuencias genómicas, el cual se enfrenta al reto de trabajar con secuencias fragmentadas, y además requiere bastante tiempo computacional. Es por esto que se desea estudiar cuáles de los algoritmos más populares existentes para el análisis de la información recolectada en el área de la metagenómica son los más eficientes, y qué métodos de optimización se le pueden aplicar.

1.2 OBJETIVOS DEL PROYECTO

1.2.1 Objetivo General

Proponer un algoritmo de agrupamiento, basado en los algoritmos que mejores resultados tengan en la clasificación de datos metagenómicos, para obtener grupos de dominio, phylum y especie de mejor calidad.

1.2.2 Objetivos Específicos

- Identificar e investigar los algoritmos que se encuentran disponibles para la comunidad enfocados en el agrupamiento de los datos metagenómicos.
- Analizar y mejorar el algoritmo de agrupamiento seleccionado para mejorar la calidad de los grupos de especies.
- Proponer los parámetros óptimos del algoritmo de agrupamiento para obtener mejores resultados en los grupos de clasificación de dominio, phylum y especie en el problema de metagenómica.

1.3 MARCO DE REFERENCIA

En las últimas décadas la metagenómica, área que se encarga del análisis de los segmentos generados en la secuenciación de ADN, ha avanzado rápidamente, permitiendo el análisis de muestras que contienen gran variedad de microrganismos y la comparación de las secuencias de ADN encontradas contra las extensas bases de datos biológicas. Aún con los grandes desarrollos que se han obtenido, la metagenómica enfrenta dos situaciones complejas que se desean solucionar; una de estas situaciones

es lograr determinar todos los organismos que existen en la muestra estudiada, y la segunda de estas es poder ensamblar porciones del genoma perteneciente a una misma especie. Esto propone un reto en técnicas computacionales para el análisis de este tipo de datos (Kelley & Salzberg, 2010).

La identificación de los organismos a los que pertenecen las secuencias obtenidas de las muestras metagenómicas se pueden realizar por medio de algoritmos de clasificación, los cuales categorizan los organismos en base al conocimiento previamente obtenido, es por esto que utiliza métodos de aprendizaje supervisados. La otra técnica de identificación es por medio de algoritmos de clusterización, los cuales son útiles para grandes poblaciones y no requieren entrenamiento con datos previos, es por esto que se conocen como métodos de aprendizaje no supervisados. A su vez estas dos metodologías utilizan un conjunto de atributos que son tomados de las muestras estudiadas, estos pueden basarse en la similitud que existen en las secuencias de ADN encontradas, o en la composición de los genomas que las componen.

Uno de los programas desarrollados con un algoritmo de clusterización basado en los componentes de los genomas es SCIMM, desarrollado por Kelley y Salzberg. Este método implementa un cambio en las cadenas de Markov que adapta la complejidad del modelo teniendo en cuenta que las muestras tomadas para el entrenamiento de este varían en cantidad. Luego se optimiza la clusterización con el algoritmo k-means. Se realizan pruebas al programa con organismos tomados de la base de datos GenBank, y por esta razón se aclara que no se puede asegurar que las muestras estudiadas sean similares a las encontradas en el medio, se debe recordar que no todos los microrganismos que existen en el medio se encuentran en las bases de datos.

Este método presenta resultados superiores a los demás métodos con los cuales lo comparan; uno de los estándares utilizados es el índice Rand, el cual es una proporción de parejas de puntos que se encuentran correctamente ubicados en un mismo conjunto o de forma separada. Este índice permite ajustar el tamaño de los clústeres. En este rango SCIMM logró un 93%, el cual es el valor más alto obtenido según Kelley y Salzberg, cuando se realiza con un conjunto de diez genomas. A medida que aumentan la cantidad de genomas disminuye la precisión del método. Se calcula que tienen un porcentaje de genomas con errores en la secuenciación del 0.5% al 2% (Kelley & Salzberg, 2010).

El software CS-SCORE también utiliza el k-means para hacer un filtro previo que busca remover los fragmentos de las secuencias que pertenecen al genoma de donde se tomó la muestra; Estos fragmentos actúan como contaminantes a la hora de realizar el análisis. En el caso de M.M. Haque se busca separar las secuencias de genoma humano y las secuencias provenientes de organismos procariota, logrando crear clústeres que se encuentran espacialmente separados. Esto se debe a que la composición de los dos genomas es muy diferente, es por esto que usan como atributo las frecuencias de tetra-nucleótidos.

Luego de realizar el filtro se procede a ejecutar un mapeo de lecturas con ayuda del algoritmo BWA fastmap. En esta etapa se busca evaluar los resultados del filtro aplicado con esta herramienta. Este proyecto obtuvo resultados positivos logrando eliminar las secuencias que se consideraban estaban contaminando la muestra, y además logro disminuir el tiempo de procesamiento y el espacio de memoria utilizados. Además este

estudio encontró que los errores en las secuencias de entrada y su longitud no alteran los resultados de CS-SCORE (Mohammed Monzoorul Haque, 2015).

En el Articulo “A new unsupervised binning method for metagenomic dataset with automated estimation of number of species“, se resalta la importancia de conocer el número de especies que contiene la base de datos para los algoritmos de agrupación no supervisados. Se propone un algoritmo “imporved fuzzy c-means (iFCM)” o un c-means borroso mejorado, el cual implementa el método varias veces con diferentes cantidades de clústeres, y luego se selecciona aquel con mejores resultados (Yun Liu, 2015)

En el trabajo realizado por Widerman Montoya se realiza una exploración de diferentes algoritmos de inteligencia artificial para la clusterización de las secuencias meta genómicas. Además se implementa un k-means iterativo que produce clústeres más puros que el algoritmo k-means. Esta nueva versión busca reagrupar los conjuntos que tuvieron una mayor distancia promedio de los demás clústeres (Montoya, 2014).

2. METODOLOGÍA

Se comenzó recolectando información y trabajos escritos por expertos, estos fueron tomados de varias bases de datos científicas accedidas por medio de internet. Se utilizó como precedente para la realización de este proyecto la tesis “Exploración Y Comparación De Métodos De Inteligencia Artificial Para La Clasificación Taxonómica En Análisis Metagenómicos” realizada por Widerman Stid Montoya egresado de Ingeniería Informática de la Universidad EIA.

Para el desarrollo del proyecto se utilizó como base el software WEKA que contiene una colección de diferentes algoritmos de aprendizaje computacional, éste es de código libre y fue desarrollado por la Universidad de Waikato. Al ser un software libre éste le permite al usuario aplicar los algoritmos implementados en el programa desde un código escrito en el lenguaje de programación Java, o inclusive, le permite al usuario modificar estos procesos que se han desarrollado y crear su propio método de aprendizaje. En el caso de este proyecto se implementa un método de aprendizaje no supervisado basado en el conocido k-means.

Además se utilizaron los servidores y las bases de datos facilitadas por el Centro Nacional de Secuenciación Genómica para realizar las pruebas del código. También se usó el software QlikView para ayudar en la visualización y el análisis de los resultados.

3. DESARROLLO DEL PROYECTO

Se comenzó el proyecto realizando una revisión de los trabajos presentados por expertos en el tema y tomando la tesis “Exploración Y Comparación De Métodos De Inteligencia Artificial Para La Clasificación Taxonómica En Análisis Metagenómicos” de Widerman Montoya.

Por los buenos resultados obtenidos con el k-means iterativo, se toma este método como base. Se llega a la conclusión de evaluar los resultados del programa aplicando la distancia coseno y euclidiana. Además se realizan pruebas para determinar el mejor valor de corte para continuar las iteraciones del algoritmo, basado en las distancias inter e interclúster.

3.1 DATOS

Se tomó como la base de datos principal una base de datos (data2_30000.fasta) proporcionada por el Centro Nacional de Secuenciación Genómica (CNSG). Esta base de datos contiene 872,576 segmentos de secuencias de ADN provenientes de diferentes especies. La base de datos fue procesada posteriormente y se crearon tres nuevos archivos de datos en los cuales se describen las secuencias según la clasificación de los seres vivos, esta se hace según dominio, la cual contiene las características más generales en la taxonomía, luego se crea una base con phylum, y por último se clasifican por especie, la cual vendría a ser en nivel más específico.

El propósito de dividir los datos según dominio, phylum y especie es identificar hasta qué nivel taxonómico el algoritmo es eficiente.

Según las diferentes clasificaciones taxonómicas se crean los tres nuevos archivos de datos, en los cuales se encuentran las siguientes cantidades de casos:

Dominio	
Bacteria	980
Eucariota	622,122
Virus	72
Phylum	
Bacteroidetes	972
Actinobacteria	17

Chordata	250,200
Chikung	0
Ascomycota	920
Nematoda	68,873
Ebola	0
Arthropoda	10,168
HIV	1
Magnoliophyta	117,585
Dengue	32
Influenza	4
Especie	
Bacteroides dorei	1,195
Bifidobacterium longum	9
Bos Taurus	157,920
Chikung	0
Aspergillus fumigatu	152
Áscaris	96,530
Ebola	0
Candida parasilopsis	776
Glossina morsitans	10,785
HIV	1
Zea mays	80,617
Malus domestica	33,370
Pantholops hodgsonii	79,865

Dengue	32
Influenza	4
Manihot esculenta	3,598

Tabla 1: Composición de las bases de datos.

3.2 ATRIBUTOS

Para representar de forma más sencilla las bases de datos se crean atributos que describen cada una de las secuencias de ADN. Estos atributos son características que buscan describir los casos de una forma estándar y más corta, y así agilizar el análisis de los datos.

Los nucleótidos y los codones se utilizan ya que estos codifican el ADN y el RNA lo que permite analizar el contenido de proteínas y aminoácidos. Estos varían en cada organismo y están relacionados con la evolución de la especie.

Otro de los atributos utilizados que también está basado en los patrones de composición de ADN es el k-mer. Este representa el ADN como estructuras de longitud k. En el caso de este proyecto se utiliza k = 4, esto se debe a que es uno de los más utilizados y que mejores resultados a presentado en el medio. Este también se conoce como frecuencia de tetranucleótidos.

Yun Liu et al realizan un estudio entre las diferentes longitudes de k-mer y llegan a la conclusión que el más efectivo para caracterizar los datos meta genómicos para la clusterización es cuando se utiliza una longitud, es decir un k, igual a 4 (Yun Liu, 2015). Un ejemplo que muestra claramente la utilidad de los tetra nucleótidos como atributo es CS-SCORE, programa que usa esta característica para crear clústeres independientes entre secuencias del genoma humano y de organismos procariotas. Los autores especifican que los fragmentos de tetra nucleótidos son valiosos ya que demuestran las diferencias en la composición de los genomas (Mohammed Monzoorul Haque, 2015).

El último atributo utilizado es el contenido de Guanina y de Citosina. Este se utiliza comúnmente para clasificar organismos en la taxonomía. Este se calcula como la suma de los pares de guanina y citosina sobre la suma de todos los pares de bases. Además como menciona Mahamuda en su artículo “Application of Machine Learning Algorithms for Binning Metagenomic Data” el contenido de GC es el atributo más importante ya que existe un enlace más fuerte entre estas dos bases que entre la Adenina y la Tiamina (Mahamuda, U, & Rasheed, 2010)

La selección de los atributos estuvo basada en opinión biológica de los expertos del Centro Nacional de Secuenciación Genómica, en los estudios previos realizados y en autores anteriores.

3.3 MÉTODOS UTILIZADOS

3.3.1 K-means

El k-means crea k conjuntos de datos evaluando la pertenencia de cada dato a un grupo determinado, esto se hace mediante la medida de la distancia del dato al centro del conjunto. Para este método se debe conocer con anterioridad el número de agrupaciones que se desean sean creadas, situación que no ocurre con frecuencia, y razón por la cual el usuario debe enfocar sus esfuerzos en encontrar un valor aproximado u óptimo para mejorar los resultados encontrados por el algoritmo.

A comienzos el proyecto se basa en el k-means iterativo desarrollado por Widerman Montoya. Se busca alterar el k, es decir el número de conjuntos, para las distintas iteraciones. Para la primera iteración que realiza el k-means se utiliza un estimado de la cantidad de especies diferentes que se encuentran en la base de datos, esta cantidad es mejor cuando se aproxima a la realidad o se encuentra por encima de esta. Para la segunda iteración se recalculan los conjuntos que se encuentran más cercanos, estos se identifican teniendo en cuenta si la distancia promedio que tienen en relación a los demás conjuntos es más grande que la suma de la media y de la desviación estándar de todos los clústeres.

Además se crea en el algoritmo la capacidad de estimar los clústeres con dos funciones de distancia diferentes, la distancia coseno y la euclidiana, las cuales se calculan con las siguientes funciones:

Distancia Coseno = $$dC(x,y) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n} x_i} \times \sqrt{\sum_{i=1}^{n} x_i}}$$

Distancia Euclidiana = $$dE(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

3.3.2 Calidad de clústeres

Con el fin de evaluar la calidad de los clústeres se toma como primer parámetro el nivel de pureza que tiene cada grupo, este se analiza como las instancias de cada clase que se encuentran en ese grupo sobre la cantidad de instancias totales que este contiene.

También se toma como métrica de evaluación la distancia que existe entre los clústeres creados. Se considera que mientras más alejados se encuentren unos de otros mejores serán los resultados, ya que se producen clústeres más puros.

Además se escoge como última medida para comparar la calidad de los grupos obtenidos el índice de Davies-Bouldin, el cual es una tasa entre las distancias inter y entre los clústeres; este cálculo estipula que los mejores algoritmos son aquellos que producen clústeres con un índice de Davies-Bouldin bajo, y se calcula con la siguiente ecuación:

$$DB = \frac{1}{k} \sum_{i=1}^{k} max_{j \neq i} \{D_{i,j}\}$$

Donde $D_{i,j}$ es la suma de la distancia inter clúster promedio del clúster i con la del clúster j dividido la distancia Euclidiana entre los centros de los clústeres i y j.

4. DISCUSIÓN DE RESULTADOS

Se realizaron varios experimentos, donde se tuvieron en cuenta diferentes variables. La primera variable que se tuvo en cuenta fue la distancia, se usó la función de distancia euclidiana y la coseno. Además se analizaron los resultados con las tres categorías de los microorganismos: La Especie que es la más específica, el Phylum al que pertenecen y el Dominio. Esto se hizo para saber hasta qué nivel discriminaba el algoritmo.

Otra variable importante que se tuvo en cuenta fueron los atributos que caracterizan las secuencias genómicas en el algoritmo. Aunque en los resultados obtenidos por Widerman en su trabajo de grado anterior el k-mer y el GC, fueron los parámetros con mejores resultados, se hicieron todas las pruebas, ya que estábamos en presencia de nuevos datos, con una complejidad mayor.

Además se hicieron varias pruebas con el valor de distancia entre clúster, para saber cuándo parar el algoritmo. Y por último y no menos importante, la cantidad de clúster entre la primera y la segunda iteración fue también una variable que se intentó ajustar en las diferentes corridas de los experimentos. Se utilizó k=10, k=15, k=20.

Por medio de las diferentes simulaciones realizadas se obtiene una gran cantidad de datos que son incorporados a la herramienta QlikView, la cual es un programa líder en la inteligencia de negocios. En ella los datos son organizados y tabulados de forma tal que las diferentes variables del sistema sean fácilmente visualizadas.

En primera instancia se desea analizar el comportamiento de k-mer y GC, que como se mencionó anteriormente fueron los atributos que mejor resultados dieron en el trabajo realizado por Widerman Montoya. Para realizar el análisis se utilizó la métrica estadística de la taza de ganancia de información, usando el método implementado en la plataforma Weka para la obtención de los resultados, los cuales se muestran a continuación:

Rango			
0,2447	190	TCGA	1
0,2393	294	CGAA	2
0,2365	100	ATCG	3
0,2325	164	TTCG	4
0,2306	295	CGAT	5
0,2067	84	AACG	6
0,2062	299	CGTT	7
0,1847	126	ACGA	8
0,1845	191	TCGT	9
0,1699	148	TACG	10
0,1666	298	CGTA	11

0,1647	312	CCAG	12
0,1535	288	CTGG	13

Tabla 2: ganancia de información de los atributos

Como se puede ver en los resultados las 13 primeras variables que tienen más relación con la clase son variables de 4-mer. Es importante notar que todas ellas tienen GC dentro de sus nucleótidos. Se escogieron de todas formas todos los atributos correspondientes al k-mer ya que esta es una base de datos representativa, y podría ser que solo este clasificando los organismos con estas características.

Se procede a analizar los resultados que el código produce para las secuencias de los genomas caracterizados desde 4-mer y GC. En la siguiente tabla se recopilan los resultados tablados según el promedio de pureza que tiene la clase que domina el contenido de instancias en el clúster y luego se pone la cantidad de clases diferentes que se encuentran en el clúster. Estos resultados se muestran para las simulaciones realizadas con 15 clústeres para las distancias coseno y euclidiana y para ambos atributos.

Clúster	GC_Cos		4mer_Cos	
	Promedio de la Máxima Clase	Cantidad de Clases en el Clúster	Promedio de la Máxima Clase	Cantidad de Clases en el Clúster
1	0,481344075	7	0,62758	10
2	0,765469062	7	1	1
3	1	1	0,999922	2
4	0,997850403	4	0,999963	2
5	0,677685654	9	0,726499	7
6	1	1	1	1
7	0,392861866	8	0,756121	9
8	0,999237603	5	0,989904	5
9	0,536809836	8	0,834393	7
10	0,998953826	6	1	1
11	0,941241558	8	0,995912	5
12	0,643215305	8	0,543246	7
13	0,956352055	3	0,999546	2
14	0,970948905	3	0,99618	7
15	0,5673231	10	0,632769	9

Tabla 3: Recopilación de resultados para la función coseno

Clúster	GC_Euc		4mer_Euc	
	Promedio de la Máxima Clase	Cantidad de Clases en el Clúster	Promedio de la Máxima Clase	Cantidad de Clases en el Clúster
1	0,995681735	4	0,997344	3
2	0,995948198	5	0,979447	6

3	0,993643825	8	0,995289	5
4	0,999213128	5	0,99689	8
5	0,999765883	3	0,999322	3
6	0,987326845	6	0,982023	7
7	0,99662048	6	0,988206	7
8	0,926932224	5	0,986816	4
9	0,974363889	5	0,967134	5
10	0,98672099	5	1	1
11	0,985509764	5	0,990724	5
12	0,984496238	5	0,957834	5
13	0,998940543	5	0,999966	2
14	0,893455725	7	0,988756	5
15	0,960013831	6	0,999727	3

Tabla 4: Recopilación de resultados para la función euclidiana.

	Cantidad Pureza Mayor a 0.8	Cantidad de Clases Menor a 3
GC_Cos	8	2
GC_Euc	15	0
4mer_Cos	**10**	**6**
4mer_Euc	15	2

Tabla 5: Resumen de resultados.

Luego de la tabla que recopila los resultados se presenta una tabla de resumen, la cual contiene tres columnas. La primera columna explica las características de la simulación que se realiza, tanto en distancia como en atributos. La segunda columna es la cantidad de todos los clústeres de esa simulación tienen un promedio de pureza mayor a 0.8. La última columna contiene la cantidad de clústeres que para la simulación dada tienen una cantidad de clases por clúster menor a 3, este número también está relacionado con la pureza del grupo.

Aunque en la tabla anterior se puede ver que la función de distancia que mejor resultados está dando es la Coseno, se muestra en los siguientes gráficos con la pureza que tienen los clústeres creados por cada una de las distancias. La figura 1 muestra los resultados de la función Coseno y la figura 2 los de la función Euclidiana.

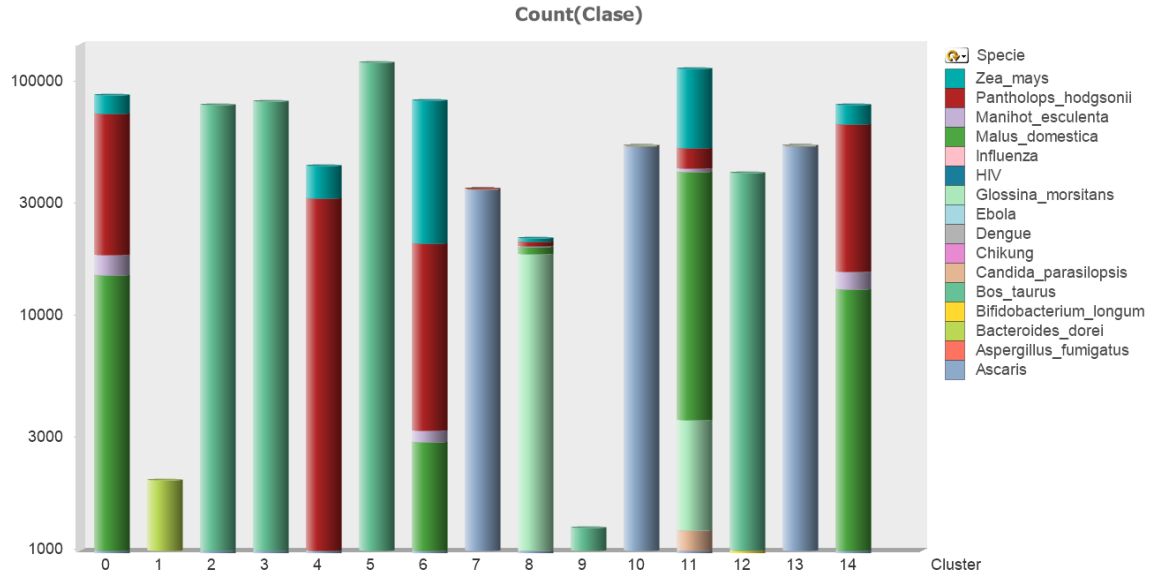

Figura 1: Clústeres creados con función coseno para la especie.

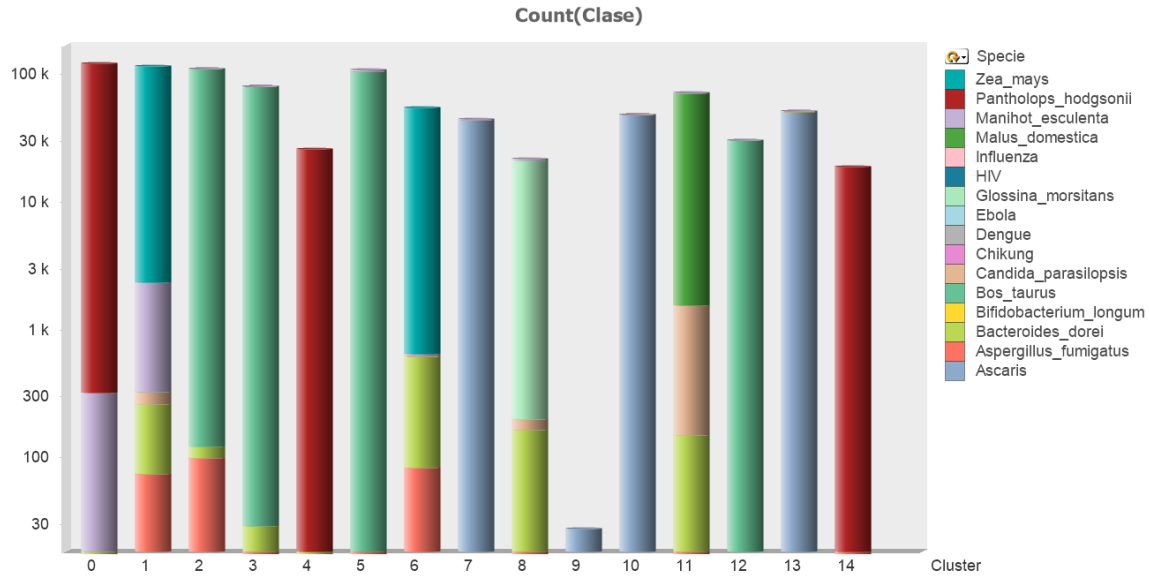

Figura 2: Clústeres creados con la función euclidiana para la especie.

En la tabla 6 se muestran las distancias promedio entre cada clúster con el resto de los clústeres. Se puede apreciar fácilmente como las distancias de la función euclidiana se encuentran más dispersas que aquellas producidas por la función coseno, además sale a la vista un valor extremo producido por esta primera función en el clúster número nueve. Este valor afecta la medida que se toma para escoger los clústeres más alejados del grupo, y es por esta razón que varios clústeres de la primera iteración no pasan directamente a la segunda y estas instancias vuelven a ser computadas.

Clúster	Promedio Coseno	Promedio Euclidiana
0	0.067	0.379
1	0.918	0.351
2	0.065	0.403
3	0.065	0.409
4	0.067	0.724
5	0.065	0.395
6	0.067	0.455
7	0.066	0.444
8	0.066	0.399
9	0.067	1.346
10	0.066	0.414
11	0.066	0.365
12	0.065	0.517
13	0.066	0.388
14	0.067	0.511
Promedio Total	0.123	0.5

Tabla 6: Distancia promedio entre clústeres.

Se desea comparar cómo se comporta el algoritmo con la clasificación de los diferentes niveles de la taxonomía. En la clasificación según el dominio se ve como separa claramente las bacterias de las eucariotas. También se puede ver que el algoritmo no separa las especies de virus y los añade a algunos de las agrupaciones de la eucariota. Es de resaltar la buena clasificación que realiza el programa con las bacterias, teniendo en cuenta además que estos genomas son más complejos.

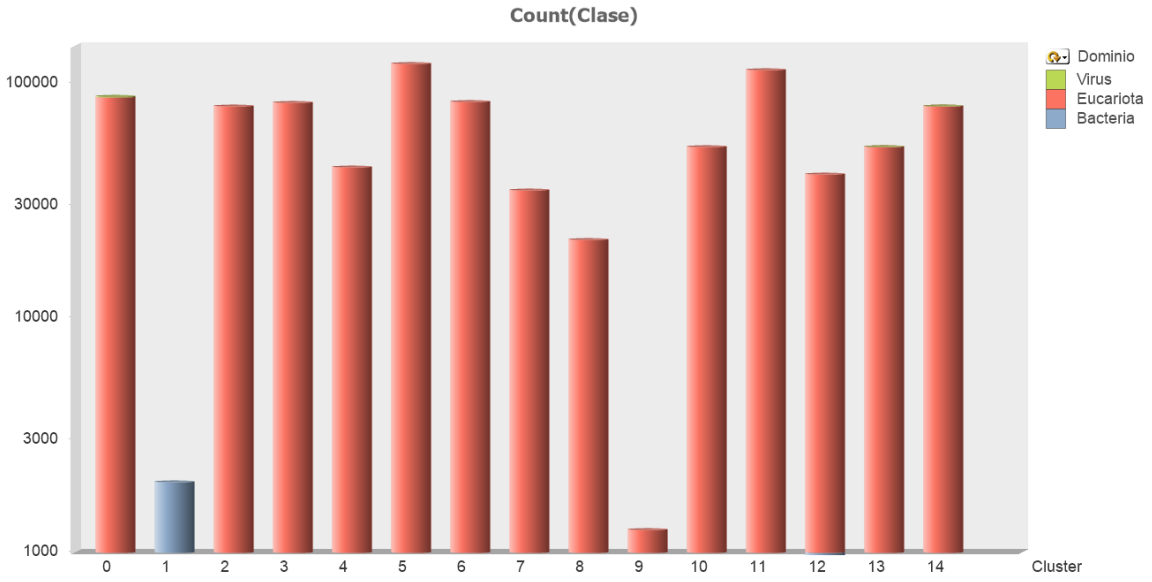

Figura 3: Clasificación según el dominio con la función coseno

En la clasificación para el phylum se observa como algunos de los clústeres siguen siendo muy puros, y como se empiezan a mezclar algunas especies. El Clúster uno es perteneciente al phylum Nematoda, el cual pertenece al dominio eucariota, en este clúster hay algunas instancias que no pertenecen a esta clase, algunas de las cuales son virus como lo vimos en la escala anterior.

También se puede apreciar que se crean varios clústeres puros de Chordata, phylum perteneciente al dominio eucariota y como en algunos de los grupos se le suman instancias de Magnoliophyta, la cual también pertenece a las eucariotas.

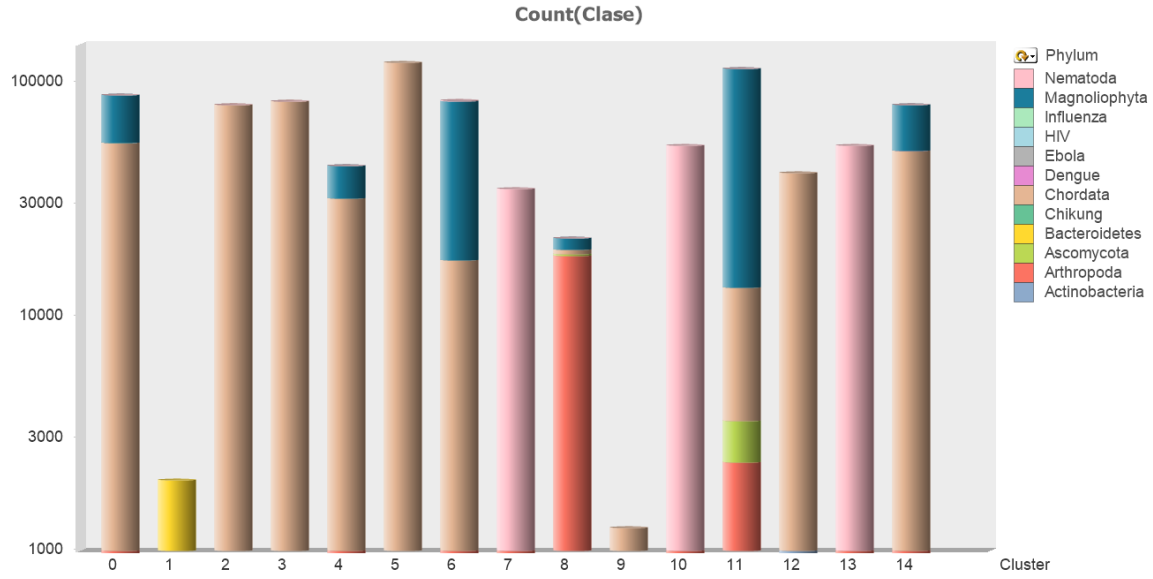

Figura 4: Clasificación según el phylum con la función coseno

Por último se analiza la especie. Se puede observar como los clústeres siguen siendo bastante buenos en su mayoría. De la especie Bos Taurus, del dominio eucariota, se crean cuatro grupos que aunque tienen unas pocas instancias de otras clases se encuentran bastante limpios. También Se puede observar cómo se logra crear un clúster con todas las instancias de Ascaris, el cual contiene muy pocas instancias diferentes a él.

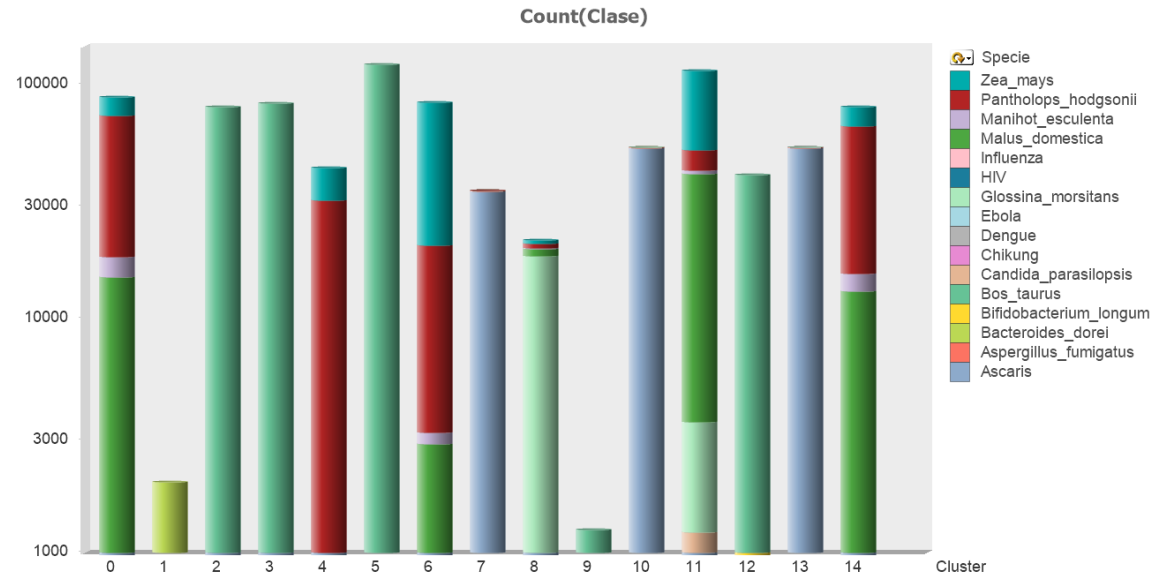

Figura 5: Clasificación según la especie con la función coseno.

Aunque no se logra separar los resultados de todas las especies, es un buen resultado poder separar las bacterias completamente de las eucariotas, Bos Taurus y Ascaris son las que logra separar mejor. Las especies Ascaris, Bos Taurus la separa, mayormente, en los clúster 2, 3, 5, 9 y 12, los cuales son bastante puros; aunque el resto de las eucariotas no las logra separar completamente. Se puede ver claramente que los virus no son separables, aunque, como se muestra en la Figura 6, aunque no los logra separar de las eucariotas no los separa tanto entre ellos, la única especie que separa en dos clústeres diferentes es al dengue.

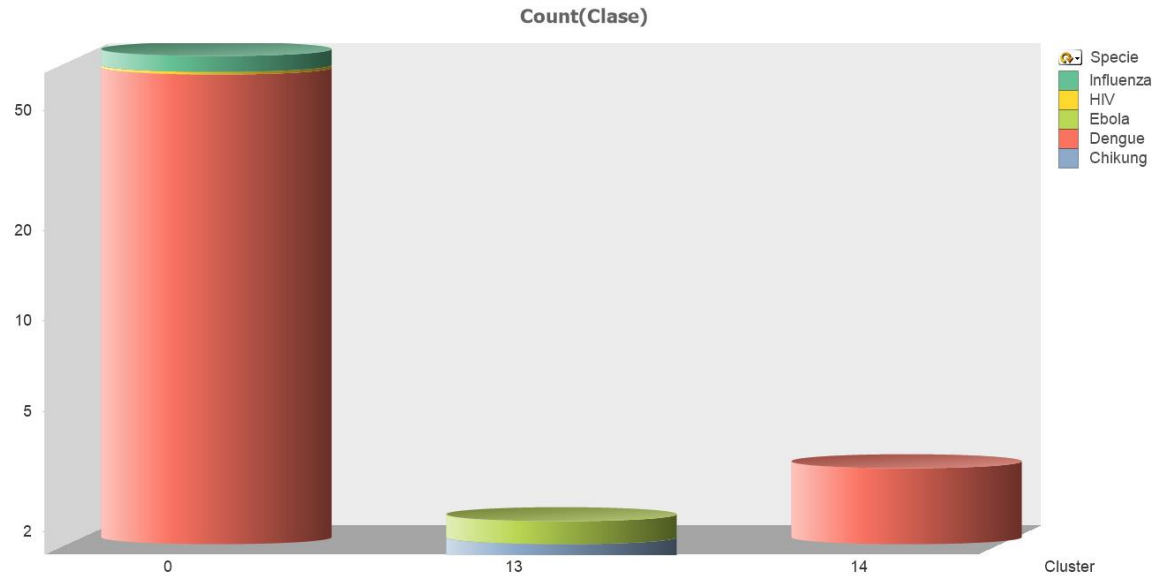

Figura 6: Clasificación de los virus.

Se procede a analizar la longitud que tienen las secuencias que se encuentran en las bases de datos con el fin de encontrar una razón para la mala clasificación de los virus. En la siguiente tabla se encuentran las diferentes especies con el tamaño de contigs más bajo que se encontró en ella y el más alto.

Nombre del Dominio	Nombre de la Especie	Tamaño Mínimo de Contigs	Tamaño Máximo de Contigs
Eucariota	Ascaris	50	30000
Eucariota	Aspergillus_fumigatus	1001	29660
Bacteria	Bacteroides_dorei	500	29906
Bacteria	Bifidobacterium_longum	540	26797
Eucariota	Bos_taurus	101	5000
Eucariota	Candida_parasilopsis	1003	29956
Virus	Chikung	11826	11826
Virus	Dengue	10392	10785
Virus	Ebola	18957	18957
Eucariota	Glossina_morsitans	101	29996
Virus	HIV	9181	9181
Virus	Influenza	853	2309
Eucariota	Malus_domestica	102	5000
Eucariota	Manihot_esculenta	1998	4998
Eucariota	Pantholops_hodgsonii	50	5000
Eucariota	Zea_mays	102	5000
	Total general	**50**	**30000**

Tabla 7: Tamaños de contigs por especie.

La figura 7 muestra los mismos resultados de la tabla anterior, pero se pueden visualizar los tamaños más facilmente. Uno de los mayores problemas para poder desarrollar un buen algoritmo de clusterización es justamente el tamaño tan diferente de estas secuencias de contigs. Pude verse la diferencia que existe dentro de un mismo organismo en sus diferentes secuencias. Otra cosa a resaltar es el tamaño de los virus. Se puede ver claramente la diferencia de tamaño de los contigs de los virus (Chikung, Dengue,Ebola, HIV,Influenza), con respecto a los demás.

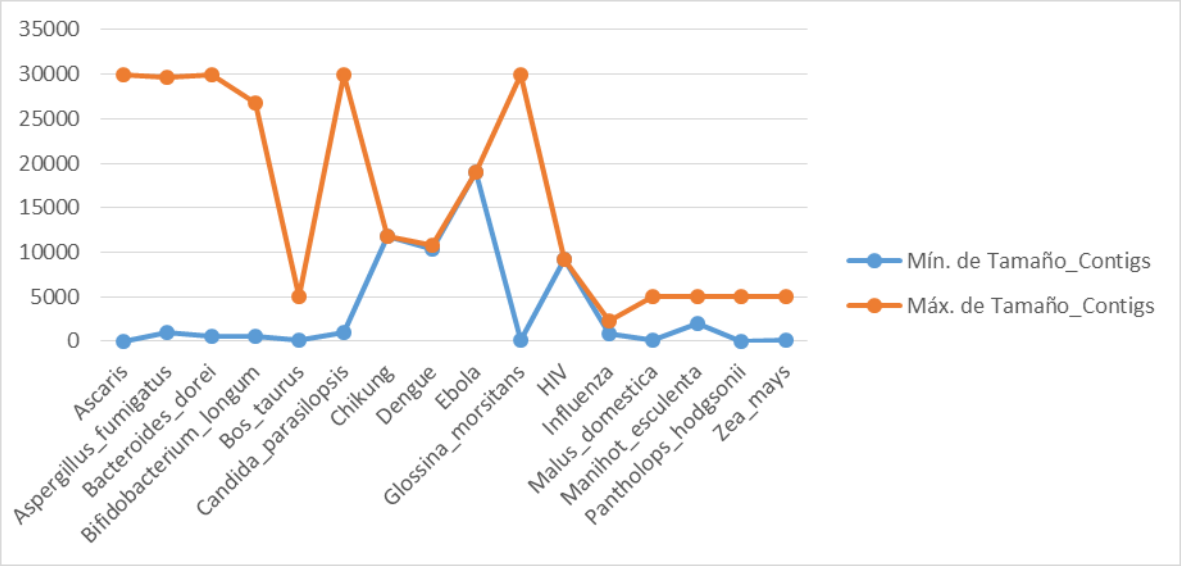

Figura 7: Gráfica tamaño de contigs por especie.

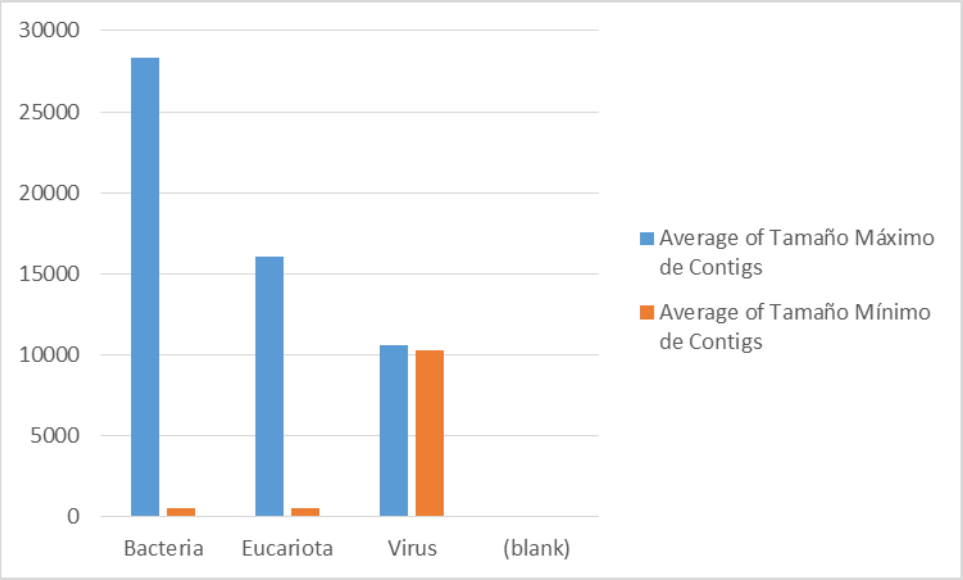

Figura 8: tamaño de contigs por dominio.

Se puede ver que los tamaños de las secuencias genómicas de los virus son en general de menor tamaño que aquellas de las otras especies.

Por último se comparan los resultados entre las dos iteraciones para evaluar el comportamiento del algoritmo. Se escogen como k para la primera iteración quince y para la segunda veinte, y el algoritmo escoge como un buen clúster para conservar en los resultados el grupo uno. También se puede observar como el algoritmo mejora algunos de los clústeres en la segunda iteración creando grupos más puros, lo que se puede

corroborar en la tabla de resumen de resultados, donde se pueden ver porcentajes de pureza muy altos y varios clústeres completamente puros.

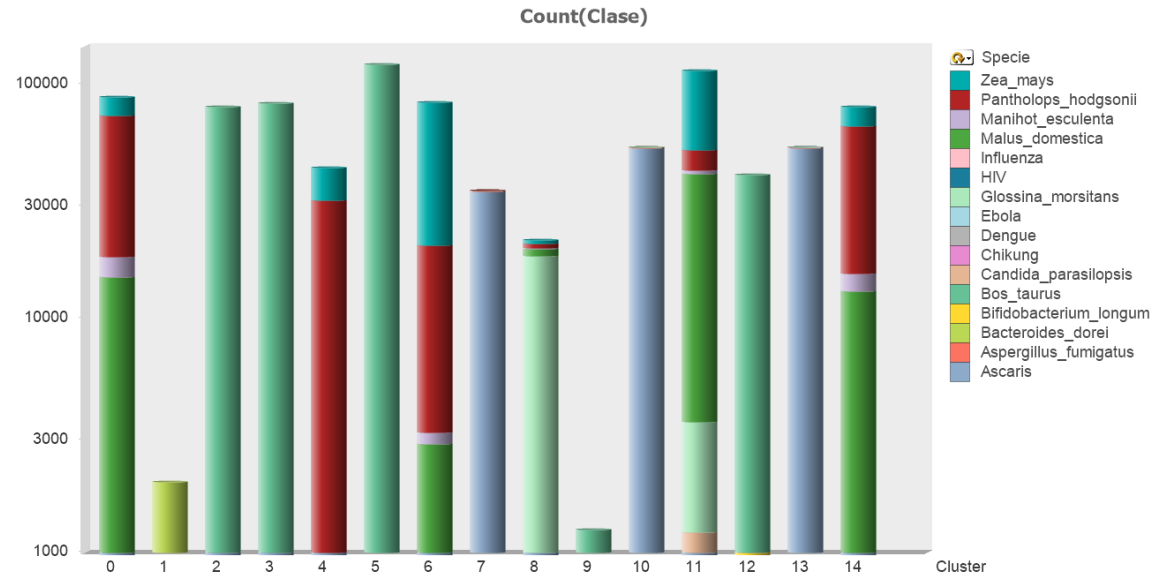

Figura 9: Primera iteración del algoritmo con función coseno y 4mer con k=15.

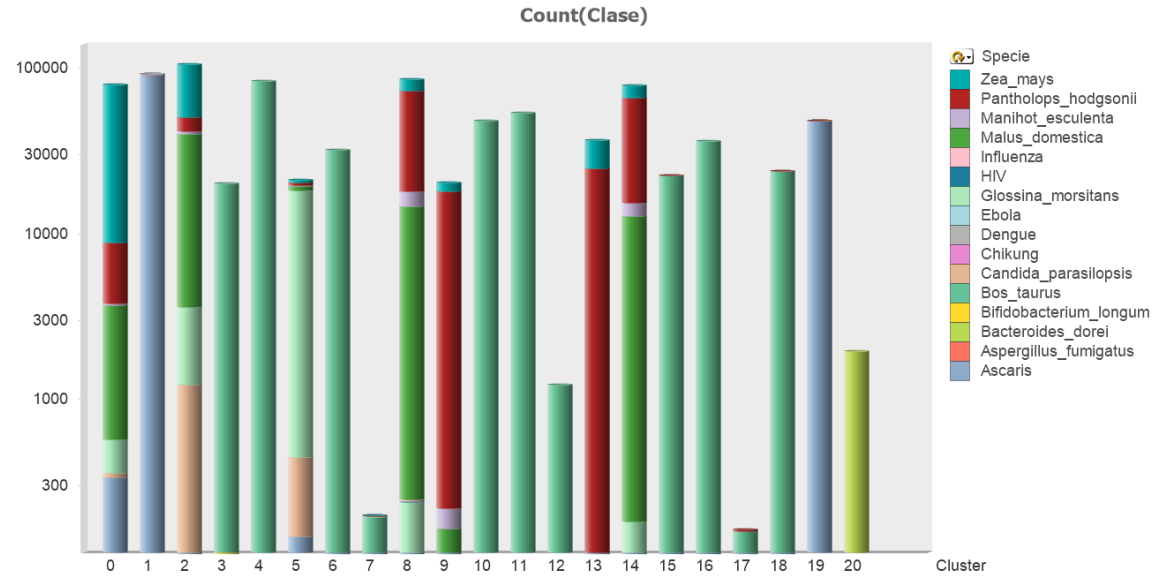

Figura 10: Segunda iteración del algoritmo con función coseno y 4mer con k=20.

	Iteración 1		Iteración 2	
Clúster	Promedio de la Máxima Clase	Cantidad de Clases en el Clúster	Promedio de la Máxima Clase	Cantidad de Clases en el Clúster

0	0,62758	10	0,887905	7
1	1	1	0,996694	6
2	0,999922	2	0,517927	7
3	0,999963	2	0,99909	2
4	0,726499	7	1	1
5	1	1	0,842179	7
6	0,756121	9	0,999905	2
7	0,989904	5	0,893401	5
8	0,834393	7	0,641757	10
9	1	1	0,877566	8
10	0,995912	5	1	1
11	0,543246	7	1	1
12	0,999546	2	1	1
13	0,99618	7	0,676769	7
14	0,632769	9	0,645444	9
15			0,383048	3
16			0,616552	2
17			0,938272	3
18			0,999185	3
19			0,9935	5
20			1	1

Tabla 7: comparación de resultados entre iteraciones.

Aunque algunos clústeres se mantienen en la segunda iteración con varios organismos sin dividir, algunos se ponen más puros que en la primera iteración, pero esto no es suficiente para decir que sea superior a la primera ya que analizando la gráfica de la segunda iteración los que logra subdividir más son los mismos organismos.

5. CONCLUSIONES Y CONSIDERACIONES FINALES

Con la realización de este proyecto se reafirma la utilidad del algoritmo k-means como herramienta para la clusterización de muestras metagenómicas, además se propone una versión iterativa que utiliza la distancia promedio entre los clústeres como método para refinar los resultados y buscar grupos más puros y exactos.

Al comparar las funciones de distancia euclidiana y coseno se identifica la función coseno como la ideal para el análisis de este tipo de segmentos genómicos, ya que en todos los niveles de la taxonomía se obtienen con ella los mejores resultados.

Se reconoce que utilizar las distancias y las relaciones entre ellas como parámetro para la evaluación de los clústeres es un buen método de calificación ya que logro identificar los grupos más alejados del resto y por ende los que se encontraban más puros. Es importante tener en cuenta los valores extremos a la hora de evaluar las distancias, ya que estos valores afectaron el margen de evaluación de los grupos lo cual se ve claramente reflejado en los resultados.

Dentro de los resultados más importantes con este algoritmo se encuentra la separación de las bacterias de las eucariotas y los virus. Dentro de las Eucariotas, se logra buena separación de Ascaris y Bos tauros, utilizando k-means con k=15 en la primera iteración y k=20 en la segunda.

Como trabajos para realizar en el futuro se identifica como una de las mayores limitaciones del algoritmo el tiempo de ejecución, se sugiere una implementación en paralelo que acelere la obtención de resultados y permita aumentar el número de grupos y de iteraciones realizadas. El reducir el tiempo que toma obtener resultados permite a los usuarios analizas más muestras con una óptima utilización de los recursos.

Por último se propone el uso de herramientas de minería de datos y Big Data para el análisis y la comprensión de los resultados que se obtienen con el programa, esto teniendo en cuenta la gran cantidad de datos que se obtienen de bases de datos robustas y que no se encuentran supervisadas.

BIBLIOGRAFÍA

Diaz, N. N., Krause, L., Goesmann, A., Niehaus, K., & Nattkemper, T. W. (2009). TACOA – Taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics, 10*, 56-56. doi: 10.1186/1471-2105-10-56

Folino, G., Gori, F., Jetten, M. M., & Marchiori, E. (2009). Evidence-Based Clustering of Reads and Taxonomic Analysis of Metagenomic Data. In V. Kadirkamanathan, G. Sanguinetti, M. Girolami, M. Niranjan & J. Noirel (Eds.), *Pattern Recognition in Bioinformatics* (Vol. 5780, pp. 102-112): Springer Berlin Heidelberg.

Kelley, D., & Salzberg, S. (2010). Clustering metagenomic sequences with interpolated Markov models. *BMC Bioinformatics, 11*(1), 544.

Kislyuk, A., Bhatnagar, S., Dushoff, J., & Weitz, J. (2009). Unsupervised statistical clustering of environmental shotgun sequences. *BMC Bioinformatics, 10*(1), 316.

Li, W., Wooley, J. C., & Godzik, A. (2008). Probing Metagenomics by Rapid Cluster Analysis of Very Large Datasets. *PLoS ONE, 3*(10), e3375. doi: 10.1371/journal.pone.0003375

Mahamuda, V., U, M. C., & Rasheed, K. (2010). *Application of Machine Learning Algorithms for Binning Metagenomic Data*. Paper presented at the Bioinformatics & Computational Biology.

McHardy, A. C., Martin, H. G., Tsirigos, A., Hugenholtz, P., & Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. [10.1038/nmeth976]. *Nat Meth, 4*(1), 63-72. doi: http://www.nature.com/nmeth/journal/v4/n1/suppinfo/nmeth976_S1.html

Mohammed Monzoorul Haque, T. B., Anirban Dutta,Chennareddy Venkata Siva Kumar Reddy, Sharmila S. Mande. (2015). CS-SCORE: Rapid identification and removal of human genomecontaminants from metagenomic datasets. *Elsevier 106*(2), 116-121.

Montoya, W. S. (2014). Exploración y Comparación de Métodos de Inteligencia Artificial Para La Clasificación Taxonómica En Análisis Metagenómicos. In I. Bonet (Ed.).

Prabhakara, S., & Acharya, R. (2011). A Two-Way Bayesian Mixture Model for Clustering in Metagenomics. In M. Loog, L. Wessels, M. T. Reinders & D. de Ridder (Eds.), *Pattern Recognition in Bioinformatics* (Vol. 7036, pp. 25-36): Springer Berlin Heidelberg.

Rosen, G. L., Reichenberger, E., & Rosenfeld, A. (2010). NBC: The Naïve Bayes Classification Tool Webserver for Taxonomic Classification of Metagenomic Reads. *Bioinformatics*. doi: 10.1093/bioinformatics/btq619

Wooley JC, G. A., Friedberg I. (2010). A Primer on Metagenomics *PLoS Comput Biol, 6*(2).

Wu, Y.-W., & Ye, Y. (2010). A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using l-Tuples. In B. Berger (Ed.), *Research in Computational Molecular Biology* (Vol. 6044, pp. 535-549): Springer Berlin Heidelberg.

Yun Liu, T. H., Liu Fu. (2015). A new unsupervised binning method for metagenomic dataset with automated estimation of number of species. Retrieved from

ANEXO 1: TABLA DE RECOPILACIÓN BIBLIOGRÁFICA

Reference	Type	Tool	Method	Features	DB
(McHardy et al., 2007)	Supervised	Phylopythia	SVM SOM	Composition based classifier Oligonucleotide composition, variable length, genome fragments	
(Li et al., 2008)	No supervised	CD-HIT	Neighbor joining	Hierarchical clustering Ultra-fast protein sequence clustering Similarity	NCBI NR (jan-07), PDB, Pfam, COG
(Folino, Gori, Jetten, & Marchiori, 2009)	No supervised	(LWproxy, GWproxy) EGWproxy	Fast heuristic algorithm	Weighted proteins	NR[3]
(Kislyuk et al., 2009)	No supervised	Markov Chain Monte Carlo	K mer (k=1, 2, 3, 4, 5)	Similarity	
(Diaz et al., 2009)	Supervised	TACOA	k-nearest neighbor	GC-content of short oligonucleotides	
(Mahamuda et al., 2010)	Supervised	Weka	Decision trees, decision tables, neural networks, support	Attributes. GC content, number of ORFs, uni-base frequency, di-base	Comprehensive Microbial Resource

			vector machine, Naïve Bayes, Bayesian networks	frequency, average length of ORF	
(Kelley & Salzberg, 2010)	No supervised	SCIMM and PHYSCIMM	Interpolated Markov model And k means	Composition-based, oligonucleotide frequencies	
(Rosen, Reichenberger, & Rosenfeld, 2010)	Supervised	NBC	Naïve Bayes	Composition based, taxonomic content	Biogas reactor dataset (Schuter 2008)
(Wu & Ye, 2010)	No supervised	AbundanceBin	Expectation-Maximization	l-tuple content, abundance of genome sequence	AMD dataset
(Prabhakara & Acharya, 2011)	No supervised	BayesianMixture	Bayesian networks	Amount of bases (Poisson distribution)	
(Yun Liu, 2015)	No supervised	MetaBin2.0	Improved fuzzy c-mean (iFCM)	k-mer frequencies	
(Mohammed Monzoorul Haque, 2015)	No supervised	CS-SCORE	k-means, read-mapping.	Genomic fragments tetra-nucleotide (4-mer)	http://dx.doi.org/10.1016/j.ygeno.2015.04.005.

Anexo 1. Tabla de recopilación bibliográfica.